基于边缘加速节点的直线加速器束流轨道参数预测技术研究

侯明阳[1,2]，郭玉辉[2]，杨旭辉[2]，杨贵进[1,*]

（1. 西北师范大学 物理与电子工程学院，甘肃 兰州 730070；
2. 中国科学院近代物理研究所，甘肃 兰州 730000）

摘要：目前能源紧缺的国际背景下，核能发展为重要清洁能源，质子加速器成为处理核废料的重要设备。在加速器束流轨道校正过程中，需要束流轨道参数辅助计算。神经网络目前应用在工业界各个领域，可以实现准确率高的数据拟合预测。因此，提出了一种基于边缘智能计算节点的直线加速器束流轨道参数预测技术。该方法通过 BPNN 对过往数据学习，生成模型部署到边缘计算节点加速预测 BPM 位置参数，可以使用边缘计算节点加速校正束流位置。实验结果表明，FPGA 硬件加速器作为边缘加速节点仿真推理速度达到 55us，能效比相比较 GPU 和 CPU 分别是其 23.13 倍倍和 553.15 倍左右，预测结果误差平均 0.5%，时延和精度达到预期目标。

关键词: FPGA；轨道参数预测；神经网络加速器；BPM

中图分类号： TL53　　文献标志码：A　　　文章编号:

0 引言

为了解决化石能源枯竭和核废料处置的问题，加速器次临界反应堆系统 ADS（Accelerator Driven Sub-critical System）被提出[1]。CiADS（China initiative Accelerator Driven System）是 ADS 的后续项目，利用高能质子束轰击金属靶实现核废料嬗变。在实际应用中，粒子加速器系统中存在多种因素导致轨道畸变，影响束流质量和稳定性，甚至引发故障，所以束流轨道校正在加速器运行中至关重要[2]。未来目标实现自适应补偿的束流轨道校正系统，其中束流轨道参数对校正束流轨道提供数据基础，需要设计一种实时性高、精度高并且可以和自适应补偿系统结合使用的束流位置预测系统[3]。

本文提出了一种基于边缘智能计算技术的轨道数据预测方法，快速预测 BPM(Beam Position Monitor)位置参数，为未来实现自适应补偿在线束流轨道校正系统提供束流位置校正。通过直接读取 BPM 数据和校正磁铁强度值构建深度学习模型，利用 FPGA(Field

收稿日期：年-月-日；**修回日期**：年-月-日
基金项目：
作者简介：侯明阳（1997-），男，辽宁新民人，硕士研究生，电子信息专业； E-mail:2021222475@nwnu.edu.cn
通信作者：杨贵进 E-mail: yanggj09@lzu.edu.cn

Programmable Gate Array)的并行计算能力和硬件优化技术加速边缘节点，实现高速、低功耗、低成本的束流轨道数据预测[4]。该方法结合了神经网络、边缘智能计算的优势，是一种有效的束流轨道数据预测方法。

1. 相关工作

1. 1 加速器束流轨道参数预测现状

目前国内外有多个机构和团队从事粒子加速器束流轨道参数预测方面的研究工作，如中国科学院近代物理研究所、欧洲核子研究中心 CERN 等。粒子加速器的性能和效率取决于束流在加速过程中的动力学行为，而束流动力学又受到多种因素的影响，如空间电荷效应、阻尼效应、集体效应、非线性效应等。已经提出的预测束流轨道参数的方法有多种，如理论模型法、数值模拟法、机器学习法等。

理论模型法最早可以追溯到牛顿的经典力学和麦克斯韦的电磁理论，它们都是用数学方程来描述物理现象的典范。其优点是既能反映内部规律，也能分析两个因素的相关关系，精度相对较高，适用于短、中、长期的预测。缺点是方程的建立需要做出一些假设条件，不同的假设会得到不同的方程，而且方程的解比较难得到。

数值模拟法是随着计算机技术的发展而兴起的，它可以对复杂和非线性的问题进行数值求解。其优点是可以处理复杂和非线性的问题，可以考虑多种影响因素，精度较高。缺点是需要大量的计算资源和时间，对计算机性能要求较高。

机器学习法是近年来受到广泛关注和应用的一种人工智能技术，它可以从数据中学习规律和知识，并进行预测和决策。这种方法是利用人工智能技术对历史数据进行分析和学习，建立预测模型[5]。优点是不需要建立数学模型，可以处理非线性和高维度的问题，具有强大的拟合能力和自适应能力。缺点是不能反映事物的内在联系和原理，需要大量且质量好的数据进行训练和验证。

2. 2 深度学习加速技术现状

深度学习在多个领域上表现出优于传统算法的效果，但是其参数体量大，若想将其应用到边缘设备中去，就要对深度学习模型进行压缩与加速[6]。

神经网络剪枝可以在保持精度的同时显著压缩模型，如 Song Han 等人实现了 35-49 倍的压缩率，其中剪枝贡献了 10 倍以上[6]。深鉴科技将其应用于 FPGA 上的语音识别[7]。Fujii 等人采用神经元修剪，将 VGG-11 网络层神经元减少 89.3%，准确性仍达 99%。权重参数减少后，FPGA 上的片上存储器可以高速访问权重存储器[8]。

模型量化可以将全精度数映射到有限整数空间，降低计算复杂度。Nagel 等人系统介绍了量化方法,量化加速效果取决于硬件平台和推理库[9]。Liang S 等人实现了二值神经网络，在 FPGA 上比 CPU 和 GPU 分别快 705 倍和 70 倍，但精度损失大[10]。Kuan Wang 等人提出了硬件感知自动量化，利用强化学习确定量化策略，有效降低延迟和能耗[11]。Riadh Ayachi 等人总结有数据量化、快速矩阵计算、频率优化、优化片内存储器设计、数据路径以及循环展开等优化方法[12]。本文中使用其中的数据量化算法，将神经网络量化为低位宽的定点数，降低数据存储空间并且加快推理速度。

2.3 BP 神经网络介绍

BPNN(Back propagation neural network)是一种基于误差反向传播算法训练的多层前馈神经网络，它能够通过调整网络的权值和阈值，最小化实际输出与预计输出之间的误差。其具有优良的非线性映射能力和灵活的网络结构，应用于函数拟合、模式识别和数据压缩等领域。

BPNN 由输入层、隐藏层和输出层构成，每一层包含多个神经元，计算过程分为正向传播和反向传播两个阶段。其结构如图 1 所示。

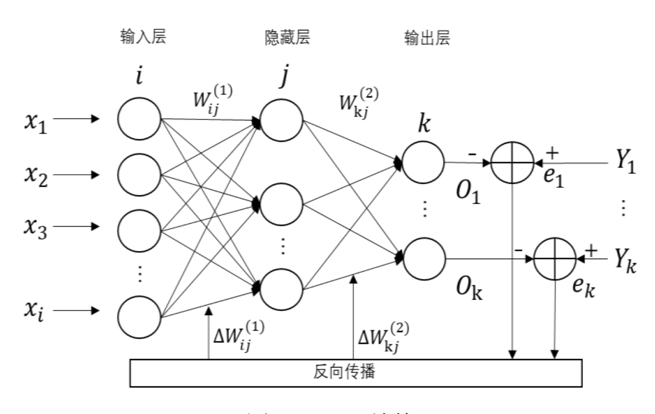


图 1 BPNN 结构

激活函数引入非线性因素，使神经网络可以拟合数据。常用的激活函数有 Sigmoid 函数，ReLU 函数以及 Tanh 函数。Sigmoid 函数和双曲正切非线性函数需要较长的训练时间，ReLU 是一个更简单易用的激活函数，如下：

$$f\left(x\right)=\max\left(x,0\right)$$

ReLU 函数在训练中收敛较快，并且具有更小的计算复杂度，所以在边缘节点部署中 ReLU 更易实现，在 HLS 中，可以进行 ReLU 函数优化。

3. 系统架构及方法设计

我们计划未来建立一套基于边缘智能计算的粒子加速器束流校正自动控制系统，作为

前期的研究，本文基于边缘智能节点实现了 MEBT 束流轨道参数的预测系统，图 2 为 BPM 以及磁铁位置关系图。边缘智能节点是指在靠近用户或数据源的网络边缘端的具有计算、存储、网络等资源的设备，其具有高效性、安全性、可移植性等优势。为了提高束流轨道参数预测的实时性和精度，使用边缘智能节点来加速执行计算和分析[13]。

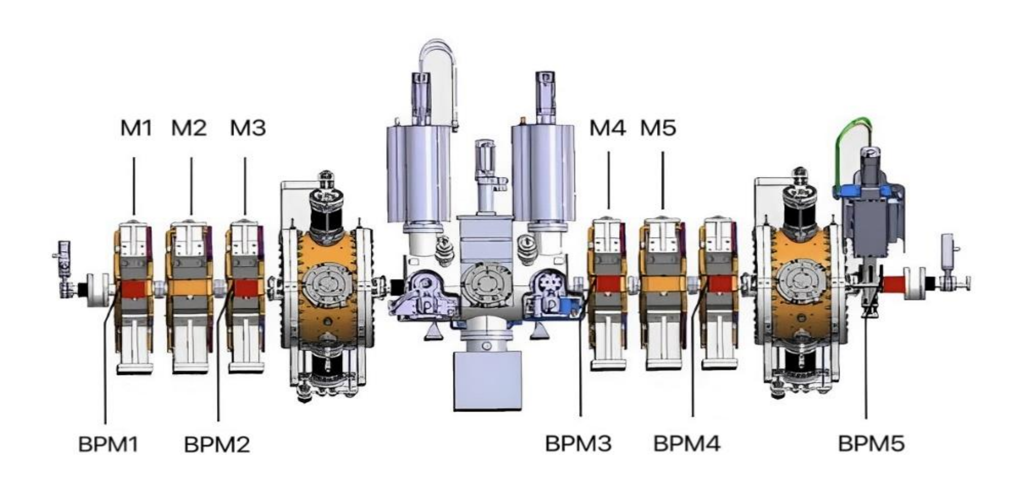


图 2 BPM 以及磁铁位置关系图

FPGA 是一种可编程逻辑器件，其可重构的逻辑单元可以实现高度并行计算，提升边缘计算节点的性能和吞吐量，并且其支持多种数据类型和算法，满足边缘计算的多样化需求。此外，FPGA 能够利用其低功耗和低延迟的特性运行算法，满足边缘计算对时效性和能效性的要求。

BP 神经网络作为一种多层前馈神经网络，能够通过反向传播算法进行自适应学习，实现复杂非线性函数的逼近。在本文中，我们利用 BP 神经网络对磁铁强度与 BPM 位置之间的关系进行函数拟合，从而预测不同磁铁强度下的 BPM 位置。BPM 位置是指粒子束在加速器中的轨道偏移量，它反映了粒子束的质量和稳定性。为了提高数据预测效率，我们将 BPNN 作为边缘节点算法部署到 FPGA 开发平台上。而 FPGA 去实现神经网络最关键的地方就是设定好一种架构，包括数据的传输、临时存储、结果读取等。神经网络的部署不只要实现权重和激活值的矩阵运算以及激活函数的输出，还要对 FPGA 架构设计进行有效优化，提高实际应用的实时性。所以设计核心是神经网络加速器设计和硬件架构设计两方面。

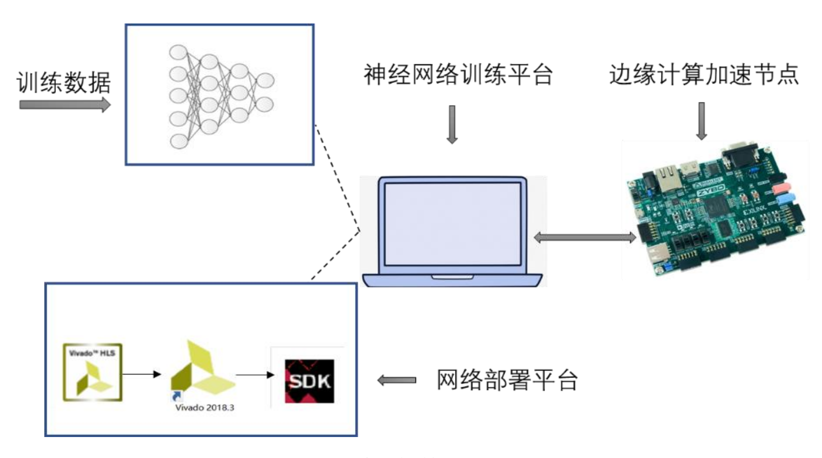


图 3 系统架构图

4、具体算法设计

4. 1 参数预测算法设计

BPNN 作为算法部署到边缘计算节点，网络体量必须要小，并且要防止输出过拟合。因此设计使用 12 个神经元输入 10 个神经元输出的单隐层 BPNN，其中隐藏层包含 20 个神经元。12 个输入神经元包括 5 个磁铁的参数值和第一个 BPM 的坐标值，10 个神经元输出包括 5 个 BPM 的坐标参数值。实验中以这些已知量为输入，预测后面 BPM 的位置坐标，这些数据坐标可以作为映射关系输出到自动化的束流校正系统，激活层采用 ReLU 函数，BPNN 网络结构设计如图 4 所示。

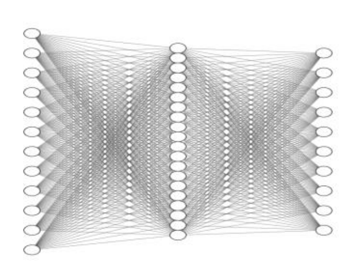


图 4 BPNN 结构设计

设计的 BPNN 包含前向传播和反向传播部分，神经网络训练首先根据输入进行正向传播即神经元输入乘权重累加偏置，从输入层到隐藏层需要进行 240 次累加乘法，从隐藏层到输出层需要进行 200 次累加乘法。反向传播使用梯度下降法，根据前向传播的各层输出结果计算 loss 函数，然后使用梯度下降进行反向传播调优，进而降低 loss 使得数据拟合。

实验的数据集是基于 Trace Win 束流动力学软件产生的模拟数据，经过统计学处理之后的数据作为训练集以及验证集，其确定性因素是 MEBT 的 lattice，如 Q 铁，非确定性因素是校正磁铁的强度，其中影响输出结果但是不可测的因素是真实加速器和模拟加速器的误差。其 BPM 数据集散点图如图 5 所示，其中每个子图的横纵坐标都分别代表其 BPM 的坐标轴数据。Magnet 磁铁电流强度参数数据散点图如图 6 所示，其中 X,Y 轴分别代表其 2 个

方向的磁铁电流强度。

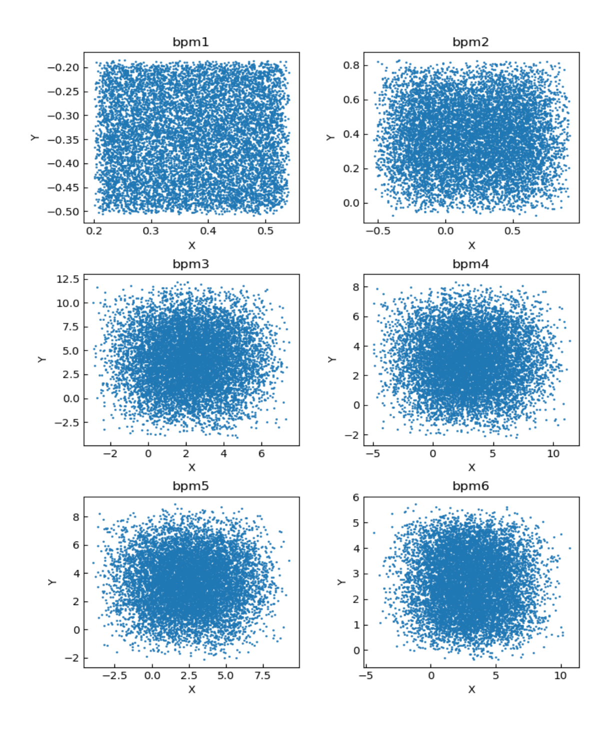


图 5 BPM 训练数据散点图

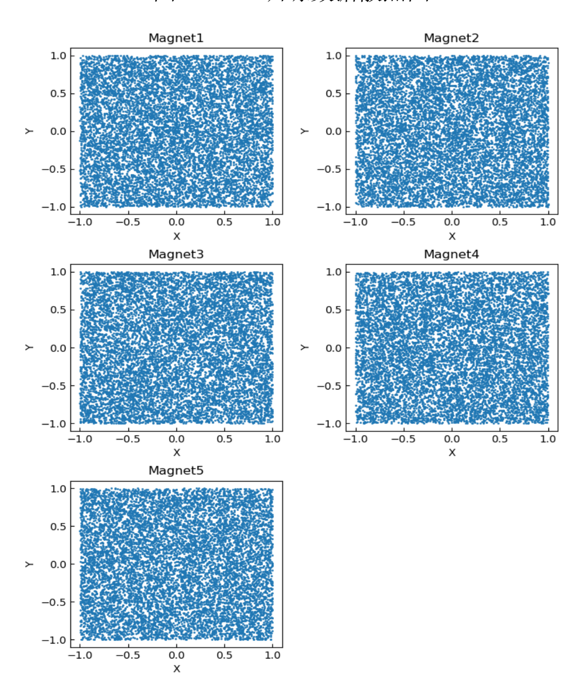
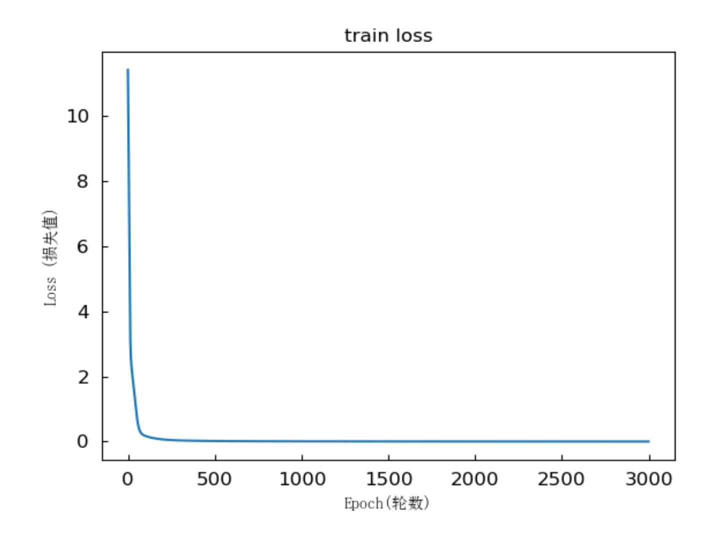


图 6 Magnet 磁铁强度参数数据散点图

训练基于 Pytorch 平台，使用随机分配的 5000 组训练集进行 3000 轮训练，BatchSize 设置为 64，对测试集和验证集进行测试，验证集误差在 0.5% 左右。神经网络训练和测试的损失值如图 7 所示，其 loss 基本收敛在 X 轴附近。



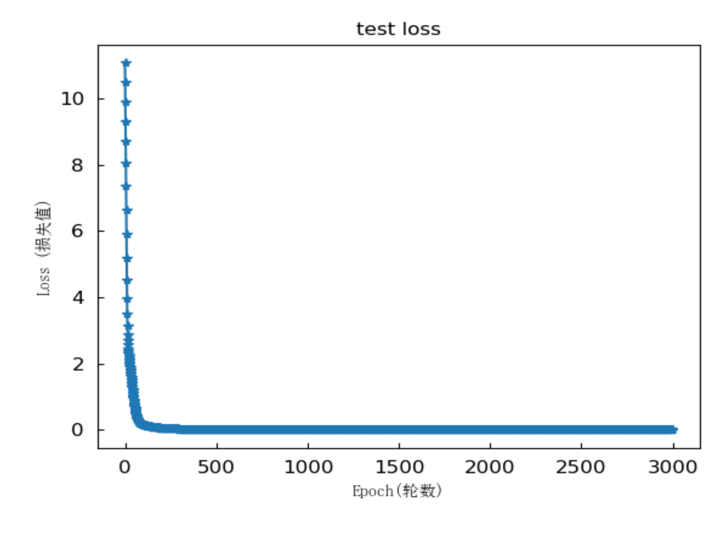


图 7 训练和测试损失值

4.2 基于 FPGA 的并行加速算法设计、实现

4.2.1 硬件系统架构设计

本文介绍了一种基于 FPGA 的神经网络加速器设计，利用 ZYNQ SOC 平台上 PS（Processing System）和 PL（Programmable Logic）之间的高效数据交互能力。神经网络加速器是一个自定义的 IP（Intellectual Property）核，它可以加速全连接层的前向推理计算，并输出预测结果。PS 端通过 AXI-lite 接口向 AXI DMA 发送控制指令。PL 端通过 AXI SmartConnect 调度数据流进行交互，使用 AXI interconnect 进行控制流的调度，DMA 的 MM2S 和 S2MM 接口分别接收数据和发送处理好的数据，完成数据流以及控制流的流通，形成闭环。图 8 是本设计的硬件系统框图。

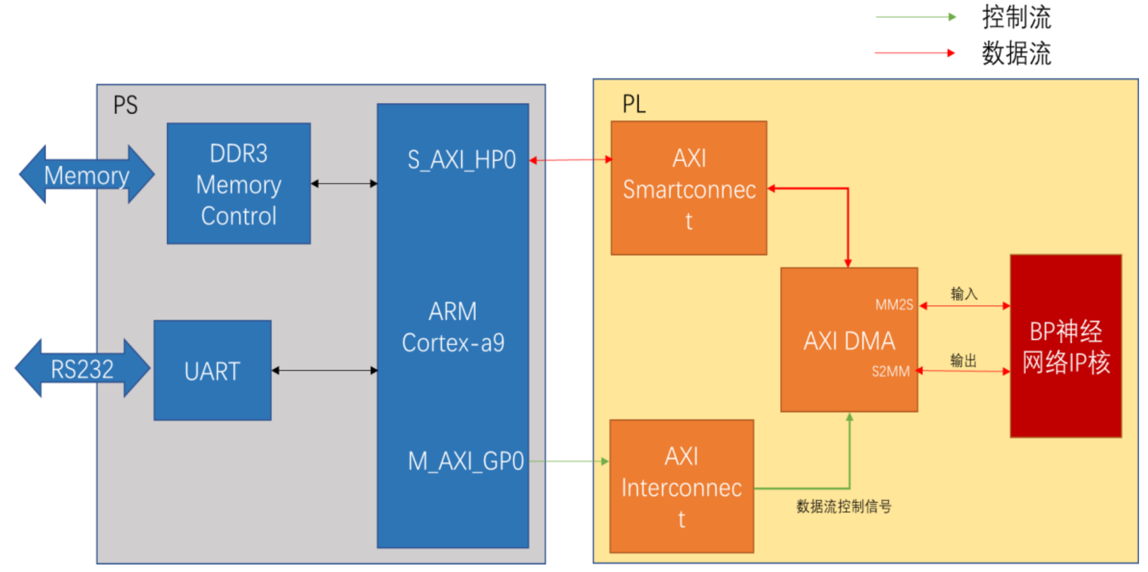


图 8 硬件系统框图

4.2.2 定点数量化算法

量化是指用较低位宽表示 32 位浮点型参数，减少推理时间和参数存储空间。神经网络的鲁棒性使量化处理参数后仍能保持精度。矩阵的浮点数乘累加运算可以通过量化数据的乘法加法和移位操作来代替,显著降低浮点数运算对 DSP（Digital Signal Processing）资源的消耗。量化主要分为训练后量化(Post-Training Quantization)和量化感知训练(Quantization Aware Training)。训练后量化不需要重新训练或标记数据，是一种轻量级的量化方法。在大多数情况下，训练后量化足以实现接近浮点精度的 8 位量化。量化感知训练需要对标记训练数据进行微调和访问，但能够实现较低精度损失的低比特量化。由于设计部署在 FPGA 平台上，定点数量化方法更适合部署，FPGA 对于定点数运算有着更高效的加速效果。

定点数量化就是将数字表示为整数和小数的组合，首先确定定点数表示的范围和精度，定点数的表示范围由位宽决定，精度则由小数点位置决定。选定合适的位宽和小数点位置，可以在同时满足精度和范围的前提下尽可能地减小所需资源。根据所选定点数的位宽和小数点位置，定义所需的数据类型和数据宽度。最后根据所选定点数的位宽和小数点位置，将浮点数转换为对应的定点数表示。

在数字信号处理和机器学习等领域中，定点量化算法被广泛应用于将高精度参数转换为低精度参数，以便于在嵌入式系统中进行高速计算时，解决高精度计算时精度不足的问题。定点量化算法可以通过不同的取整函数来将浮点数转换为定点数。向下取整会导致精度损失，而向上取整则可能会导致溢出。因此，定点量化算法需要根据应用场景和需求选择合适的取整方式和量化位数。

4.2.3 神经网络加速算法

BPNN 作为全连接神经网络，其计算本质就是权重矩阵乘法累加偏置，所以对于矩阵乘法的优化是设计的核心。本文采用流水线和循环展开的优化方法，矩阵乘法算法伪代码如图 7 所示，可以看到其中主要有 5 个循环，有两个矩阵乘计算并且累加偏置。矩阵乘法应该是 A 的每一行的元素乘 B 的每一列的相应元素的叠加。矩阵乘法的运算逻辑最内层的循环是进行 A 的第 i 行元素与 B 的第 j 列元素的乘积累加。其外层是对矩阵 B 的列遍历，可以计算出其第 i 行的所有元素，最外层的循环是对行的遍历。矩阵向量乘法被定义成：

$$c_{ij} = \sum_{k=0}^{n-1} a_{ik} b_{kj} \quad ,0 \le i, j \le n-1$$

```
for i = 0 to n − 1 do
    for j = 0 to n − 1 do
        output1 = bias1[i] + weight1[i][j] ∗ in[j];
    end for
end for
Relu(output1);
for i = 0 to n − 1 do
    for j = 0 to n − 1 do
        output = bias2[i] + weight2[i][j] ∗ output1[j];
    end for
end for
```

图 9 矩阵乘法算法伪代码

矩阵运算是计算资源的主要消耗者,对于设计的 BPNN 来说，主要加速部署的就是前向传播部分，前向传播就是由累加乘组成的，输入和输出的数据和设计的神经网络相同。根据前面 BPNN 的结构设计来看，硬件加速器应该包含从输入层到隐藏层的 240 次累加乘和从隐藏层到输出层的 200 次累加乘。为了提高循环并行性能，我们使用循环展开和流水线化技术。循环展开时将循环体复制多份，并在硬件上同时执行一组迭代。其参数指定了每次在硬件上并行执行的数目。循环展开不仅可以解决利用率低下的问题，还可以帮助优化数据路径和片上存储器设计。FPGA 流水线化是指将一个复杂的计算任务分解为若干个子任务，并按照顺序组装成一个流水线处理的过程。FPGA 部署神经网络的流水线化是利用 FPGA 上的可重构资源进行了设计和优化，使用 DSP 进行矩阵运算,使用 LUT 进行逻辑运算和查找表操作。整体硬件加速器内部架构如图 10 所示。

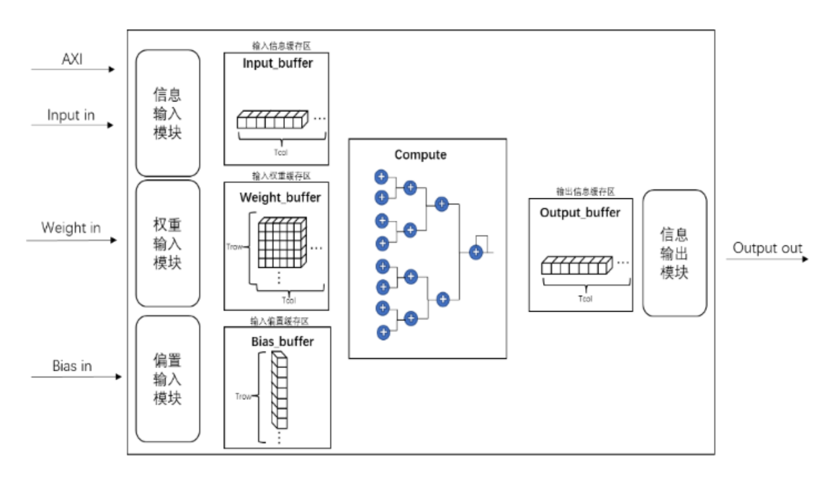


图 10 硬件加速器内部架构

5、测试及讨论

本设计应用了 BPNN 神经网络技术对于过往的实验数据进行拟合，得到期望的神经网络模型，使用定点量化技术对模型参数量化，根据硬件整体架构以及其硬件加速器的设计，将 FPGA 作为边缘加速节点部署 BPNN 进行加速。测试包括验证边缘加速节点的硬件加速效果、量化性能、硬件性能指标以及硬件系统架构的可行性。

5. 1 软硬件实验环境

本文测试部署边缘计算节点的神经网络模型为 BPNN，使用 Pytorch 训练平台对神经网络进行 7000 轮训练，将得到的神经网络权重以及偏置文件保存下来作为后续测试的材料。使用的实验开发板是 ZYBO-Z7,主频为 120MHz；CPU 对照组采用 12 vCPU Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50GHz 处理器，主频 2.5GHz，处理器数量为 24 核 48 线程，Linux 操作系统；GPU 对照组采用 Nvidia GTX 3090 芯片，拥有 10496 个 CUDA 核心，核心频率 1400MHz，配有 24GB 的显存，Linux 操作系统；在 ZYBO-Z7 开发板上对本设计进行了实验，硬件加速器的资源利用率如表 2 所示。

表 2 硬件加速器的资源利用率

类别	FF	LUT	DSP48E	BRAM_18K
资源总数	35200	17600	80	120
使用数	8039	15808	64	0
使用率	22%	89%	80%	0%

从表 2 看到加速器 BRAM 的资源利用率为 0，是因为神经网络的模型体量比较小，参数经过优化放在 LUT 中实现。本文使用 Vivado 2018.3 对设计完成综合实现。

5. 2　硬件加速器加速效果测试

本设计主要使用了矩阵计算加速、循环流水线以及循环展开技术对神经网络进行硬件加速，为了得到实际加速效果，需要对使用上述优化技术加速前后的资源占用以及时钟周

期进行测试对比，加速前指未使用流水线化、循环展开等优化技术的资源占用，加速后指使用了优化技术的资源占用。如表 1 所示，对比硬件加速器加速前后的资源占用率，我们发现 DSP48E 和 LUT 的占用率明显上升。符合硬件加速器用资源换取运行速度的设计理念，从加速前后时钟周期可以看出，硬件加速器对算法加速了 18.05 倍左右，硬件加速器仿真前向推理时间为 2.9us。

表 1 硬件加速器优化前后资源对比

类别	时钟周期	FF	LUT	DSP48E	BRAM_18K
加速前	5236	3%	10%	6%	1%
加速后	290	22%	89%	80%	0%

5.3 量化性能测试

量化参数之前需要测试不同位数定点数量化算法的误差，确定量化位数之后将浮点值转换成定点数并且对其进行 round 操作。分别进行 8 位量化、16 位量化以及 32 位量化的实验，将神经网络的权重参数通过 HLS 工具里的 AP_fixed 函数分别量化成 8 位、16 位以及 32 位，将这些不同位数的数据送入神经网络加速器中得到预测结果，对预测结果和实际值进行计算误差，得到不同位数量化的误差率，从而得到不同量化位数对输出结果的精度影响。实验结果如图 9 所示，图中各点分别代表对应 BPM 位置的量化后误差率的数值，3 条折线分别代表 3 种量化位数的精度。由图 11 可见 16 位定点数量化的误差很小，与 32 位定点量化精度基本一致， 8 位量化的精度损失明显，精度损失最大的点位已经超过了 40%，所以本设计采用 16 位的定点量化方法。8 位定点量化的误差较大原因是输入值的范围已经超过了 8 位定点数的表示范围，造成误差偏移严重。

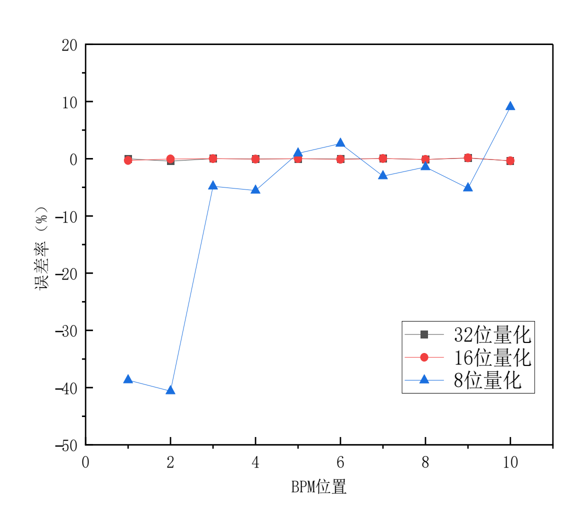


图 11 量化位数误差对比

5.4 硬件性能评估

为了测试本设计的性能指标，需要对照组进行功耗测试、有效算力和能效比的性能评

估。选取目前主流的硬件加速器作为对照组，包括 GPU 和 CPU 设备。使用包含 CPU 和 GPU 的 Linux 服务器进行测试，测试数据是 TraceWin 输出的经过归一化及统计学处理的 10000 组数据。从验证集中随机抽取 10 组，分别送入 ZYBO-Z7、GPU、CPU 计算平台，对于 GPU 和 CPU 平台使用 profile 工具进行前向推理分析，可以得到满载情况下其前向推理的时延，对每组数据在不同计算平台上的处理时间、网络推理时间和运算量及不同平台的计算功耗进行测试，测试结果如表 3 所示，FPGA 开发板功耗测试如图 12 所示。从表 3 可看出，BP 神经网络在该硬件加速器推理的平均时间约为 55us，计算速率分别是 GPU 与 CPU 的 0.36 倍、7.618 倍，说明本文设计的边缘智能节点在 FPGA 上加速效果明显，加速效果介于 GPU 和 CPU 之间。该节点实际功耗仅为 1.572W，远低于 GPU 与 CPU，适合使用于嵌入式移动设备。BPNN 网络的运算量为 440 FLOPs= 0.00044MFLOPs，故该加速器每秒的有效算力为 0.00044/0.000055= 8MOPS，能效比为 8/1.572= 5.09 MOPS/W。该加速器的能效比是 GPU 的 23.13 倍，CPU 的 553.15 倍，可以看出本文设计的边缘计算节点有着优秀的性能收益。实验结果表明，系统整体不但拥有神经网络的数据拟合能力，还结合了边缘加速节点的硬件加速优势，显著降低运行时延，增加算法实时性。

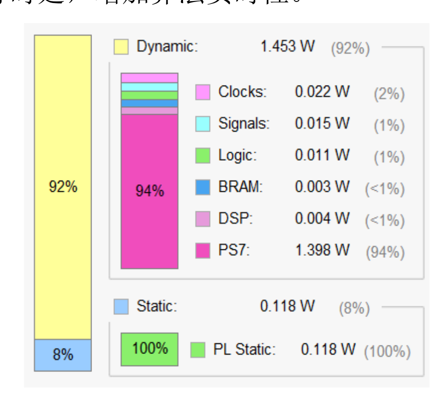


图 12 开发板功耗测试

表 3 加速平台性能测试对比

类别	ZYBO-Z7	GPU	CPU
单组数据推理时间/us	55	20	419
每秒有效算力/MOPS	8	22	1.05
实际功耗/W	1.572	102	113
能效比/MOPS/W	5.09	0.22	0.0092

5. 5 板级系统功能测试和精度测试

将设计好的硬件加速 IP 核导入到 Vivado 软件之后，根据硬件架构设计绘制 Block Design 设计图，将硬件设计集成到 BSP（Board Support Package）包中，导入 SDK 软件进行 PS 端控制代码编写。输入数据后，经过 AXI 接口传递到硬件加速器，FPGA 的 PL 端数据处理后，通过 PS 端控制并输出 BPM 位置数据，ZYBO 开发板功能串口测试结果如图 13

所示。SDK Terminal 连接 COM4 串口后设置比特率为 115200，在 SDK Terminal 的输入窗口填写 BPNN 的输入数据，点击 Send 之后数据送入开发板中，之后串口正常输出预测的 BPM 位置数据。实验表明，本文设计的硬件架构可以正确接收并处理神经网络输入数据，验证了本系统架构设计的正确性。

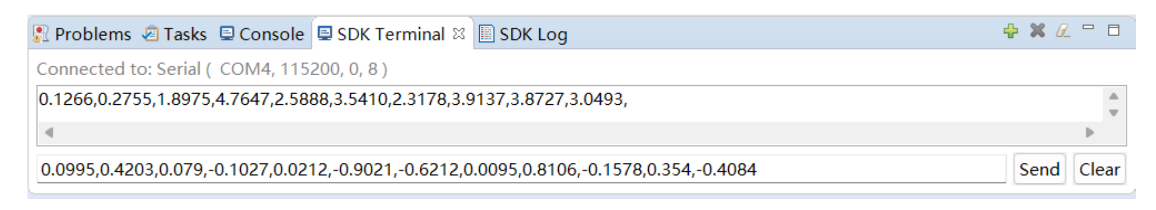


图 13 板级串口测试结果

从验证集抽取的 10 组数据输入边缘加速节点中进行精度测试，将得到的预测数据与原始数据对比，得到各组预测结果和实际结果的误差如图 14 所示，BPM 预测误差在 0.5%左右。

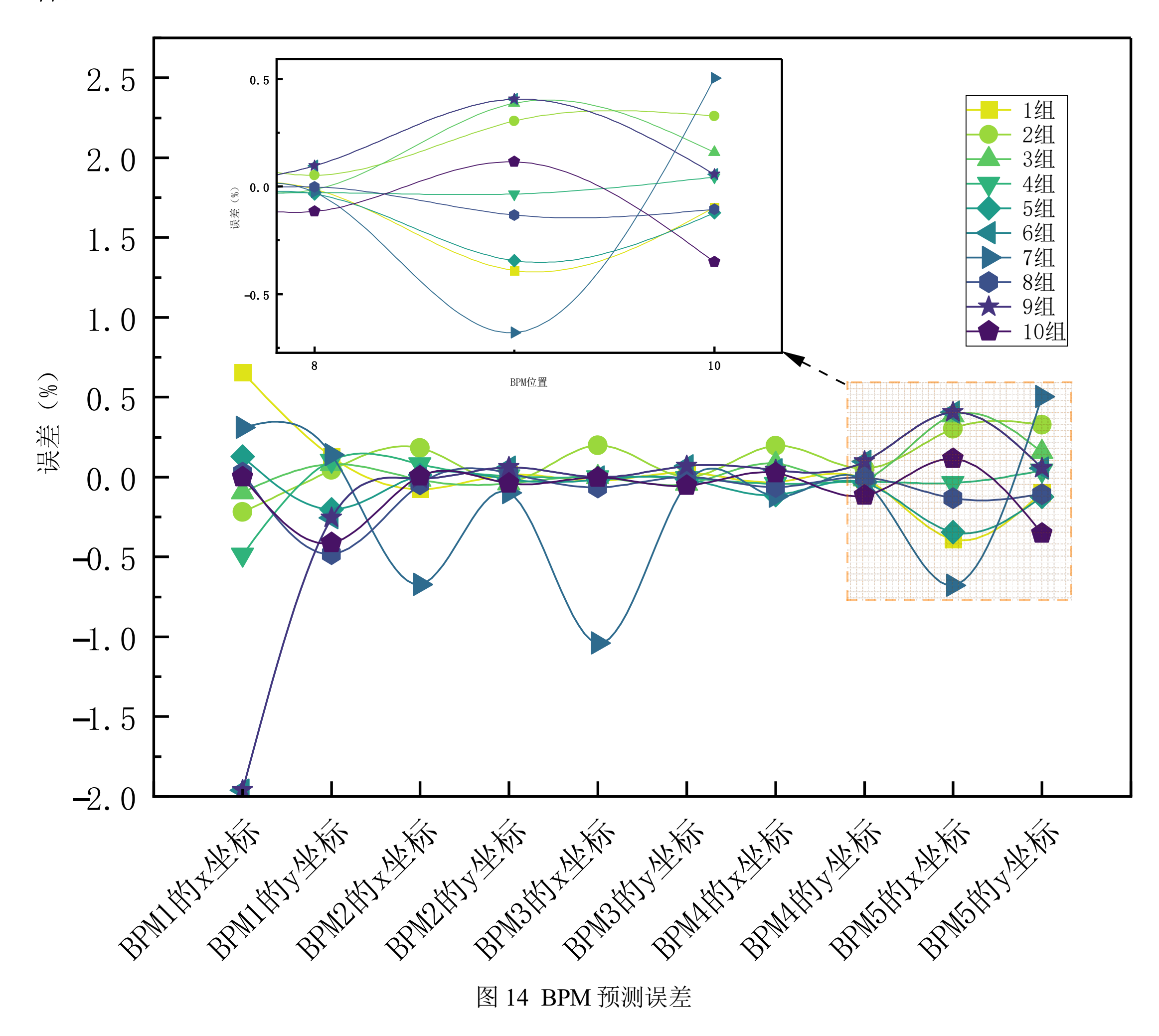


图 14 BPM 预测误差

6. 结论及展望

本文从提高计算速度和提高能效比的角度出发，设计了一种以 FPGA 作为边缘智能节点的加速系统，通过软硬协同的方式，优化神经网络前向传播电路及降低了带宽，以满足

束流轨道参数预测的实际需求。神经网络加速器在计算性能上达到8MOPS的有效算力，硬件加速器的计算功耗为1.572W，能效比为5.09MOPS/W，是GPU的23.13倍，是CPU的553.15。从数据中可以得出，作为神经网络边缘加速平台，FPGA 加速神经网络可以取得很好的加速效果，其加速效果在实验对比中效果介于 GPU、和 CPU 之间，并且明显强于 CPU。在具有良好加速效果的情况下，FPGA 加速器具有很高的能效比，在大规模部署的情况下更加节省能源。

实验结果表明，本文提出的硬件加速器适合应用于边缘计算设备，并且在对 BPM 位置预测中预测值和真实值平均误差在 0.5%，硬件加速器仿真平均推理时延 2.9us，FPGA 边缘节点推理时间平均为 55us，能够满足未来实现的自适应补偿在线束流校正系统的时延要求，这里 FPGA 边缘节点推理时间的大部分时延都浪费在了数据传输，未来目标设计更低时延的数据传输硬件架构，以达到更好的加速效果。具体而言，主要贡献为研究了神经网络加速技术在粒子加速器束流轨道数据预测中的应用，探索了一种基于边缘智能计算节点的直线加速器束流轨道参数预测技术研究，并通过实验验证了其应用的可行性。本设计的核心是边缘加速节点，具有很强的移植性，该硬件架构以及边缘节点的加速算法不仅可以进行BPNN 的前向推理加速，对于其他有高实时性要求的算法或者神经网络也可以很好的加速。但本文的研究工作是基于虚拟加速器的数据进行的，对于真实的加速器的束流轨道数据预测方案，需要考虑在实际加速器中更多的影响因素，在实际加速器中进行数据清洗，选择典型的数据进行学习。在今后的工作中，我们将对神经网络模型及加速方案进行改进，未来结合自适应补偿的自动化束流校正系统应用于真实的粒子加速器。

参考文献:

[1] YANG Xuhui . Neural Network-based Calibration Technique for C-ADS Injector II Beam Offset D]. Lanzhou University, 2019. (in Chinese).

（杨旭辉. 基于神经网络的 C-ADS InjectorII束流偏移校准技术研究[D].兰州大学, 2019.1-41 DOI:10.27204/d.cnki.glzhu.2019.000051.）

[2] Wan Jinyu, Sun Zheng, Zhang Xiang, et al. Machine learning applications in large particle accelerator facilities: review and prospects[J]. High Power Laser and Particle Beams, 2021, 33: 094001(in Chinese).

(万金宇, 孙正, 张相, 等. 机器学习在大型粒子加速器中的应用回顾与展望[J]. 强激光与粒子束, 2021, 33: 094001. doi: 10.11884/HPLPB202133.210199)

[3] Yuanshuai QIN, Zhijun WANG, Chi FENG, et al. Longitudinal Beam Parameters Measurement by Beam Position Monitors[J]. Nuclear Physics Review, 2021, 38(1): 30-37. doi: 10.11804/NuclPhysRev.38.2020055

[4] JIAO Licheng,SHUN Qigong,YANG Yuting,etal. Deep Neural Network FPGA Design Progress, Implementation and Outlook[J]. Journal of Computer Science and Technology, 2022, 45(3)(in Chinese).
(焦李成, 孙其功, 杨育婷, 等. 深度神经网络 FPGA 设计进展，实现与展望[J]. 计算机学报, 2022, 45(3). doi: 10.11897/SP.J.1016.2022.00441)

[5] Meier E, Morgan M J, Wu J. Electron beam energy stabilization using a neural network hybrid controller at the Australian Synchrotron Linac[J]. Proc. PAC'09, 2009: 1201-1203.doi: 10.1063/5.0030416

[6] Han S , Mao H , Dally W J .Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding[J].Fiber, 2015, 56(4):3--7.doi:10.48550/arXiv.1510.00149

[7] Han S , Kang J , Mao H ,et al.ESE: Efficient Speech Recognition Engine with Compressed LSTM on FPGA[J], Association for Computing Machinery. 2017.doi: 10.1145/3020078.3021745

[8] Fujii T , Sato S , Nakahara H ,et al.An FPGA Realization of a Deep Convolutional Neural Network Using a Threshold Neuron Pruning[C],International Symposium on Applied Reconfigurable Computing.Springer, Cham, 2017.doi:10.1007/978-3-319-56258-2_23

[9] Nagel M, Fournarakis M, Amjad R A, et al. A white paper on neural network quantization[J]. arXiv preprint arXiv:2106.08295, 2021. doi: 10.14711/thesis-991012980216903412

[10] Liang S, Yin S, Liu L, et al. FP-BNN: Binarized neural network on FPGA[J]. Neurocomputing, 2018, 275: 1072-1086.doi: 10.1016/j.neucom.2017.09.046

[11] Wang K, Liu Z, Lin Y, et al. Haq: Hardware-aware automated quantization with mixed precision[C],Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 8612-8620.doi: 10.1109/cvpr.2019.00881

[12] Ayachi R, Said Y, Ben Abdelali A. Optimizing neural networks for efficient FPGA implementation: A survey[J]. Archives of Computational Methods in Engineering, 2021: 1-11. doi: 10.100//s11831-021-09530-9

[13] Yang X, Chen Y, Wang J, et al. Online beam orbit correction of MEBT in CiADS based on multi-agent reinforcement learning algorithm[J]. Annals of Nuclear Energy, 2022, 179: 109346. doi: 10.1016/j.anucene.2022.109346

Research on Prediction Technology for Beamline Parameters of Linear Accelerator Based on Edge Computing Nodes

Received date:2023-05-dd; Revised date:2023-mm-dd
Foundation:

HOU MingYang[1,2], GUO Yuhui [2], YANG Xuhui [2], YANG Guijin[1,*]

(1. College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou 730070, China;
2.Institute of Modern Physics, Chinese Academy of Sciences, Lanzhou 730000, China;)

Abstracts: In light of the current international energy scarcity, nuclear power has emerged as a crucial source of clean energy. Proton accelerators have therefore become a pivotal technology in nuclear waste management. During beamline orbit correction processes, precise calculations of beamline orbit parameters are required. Given the demonstrated effectiveness of neural networks in a wide variety of industry domains, they offer promising potential for high-accuracy data fitting and prediction. Hence, this study proposes a novel direct linear accelerator beamline orbit parameter prediction technique based on edge intelligence computing nodes. This technique leverages BPNN to learn from historical data and generate a powerful model that can be seamlessly deployed to edge computing nodes, thereby accelerating the prediction of BPM location parameters. Furthermore, the proposed approach may be complemented by an adaptive compensation system in the future, which, in combination with edge computing nodes, could enable automatic beamline position correction, thereby achieving beamline orbit correction. Our experimental results demonstrate that FPGA, as an edge acceleration node, can achieve an inference speed of 2.5us, which represents a remarkable performance enhancement of approximately 165.6 times compared to CPU and approximately 7.9 times compared to GPU. The predicted results exhibit an average error of only 0.5%, and they exhibit the desired latency and accuracy characteristics.

Key words: FPGA; Prediction of orbital parameters; neural network accelerators; BPM